Coupling metagenomics with high-performance computing to mine the Sequence Read Archive (SRA) to analyze Pseudomonas phage PAK-P1

Sruthi Ganapaneni[1], Haley Leffler[1], Bhavya Papudeshi[2], Sheri A. Sanders[2], and Thomas A. Doak[2]

[1]Department of Human Biology, Indiana University; [2]National Center for Genomic Analysis Support, Pervasive Technology Institute, Indiana University

Introduction

- The **Sequence Read Archive (SRA)** is a NCBI database currently hosting ~14PB of sequence data with metadata; most published genomics projects around the world deposit their "raw" sequence data here, available for further downstream analysis.
- Thus, it is an extremely large database, and a special tool is required to "mine" the SRA
  - **Search SRA** was developed to look through subsets of sequences in SRA, which is a more manageable amount of data to search against
- **Metagenomics** is a technique to analyze all of a microbial community or microbiome, and uses high-performance computing to analyze these huge quantities of genomic data
  - **The SRA** contains many metagenomic datasets
  - **Jetstream** is a cloud computing environment that houses a variety of pre-configured Virtual Machines (VMs)
- The overall goal of this project is to develop a workflow to mine the existing Sequence Read Archive (SRA)
  - The results will be compared against datasets of interest to generate a visual representation of the relationships between the metagenomes, in order to establish patterns
- **Developing workflows** helps NCGAS in its mission to provide support and tools to researchers with little technical experience and reduce work that could take months to as little as a day
- The workflows developed in our projects will be shared with the community to help other researchers

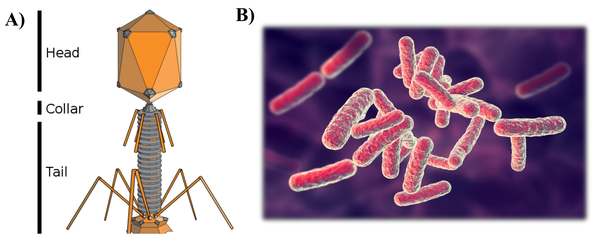


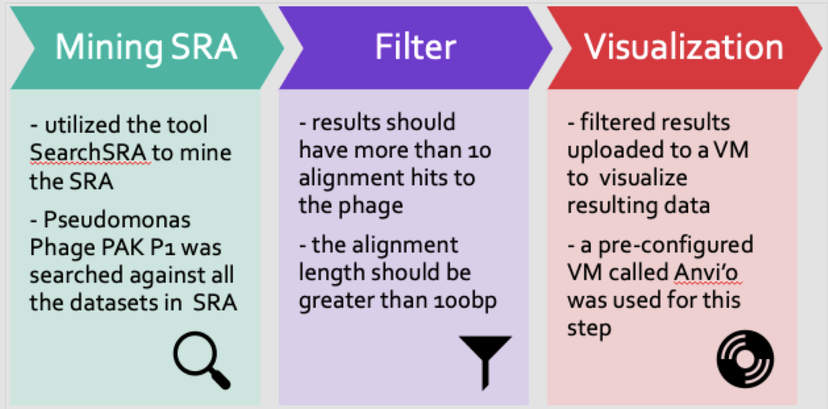
Figure 1A: Bacteriophage image, B: *Pseudomonas aeruginosa* bacteria

Case Study

- I am applying metagenomic methods to develop a microbiome analysis workflow to identify datasets that have **Pseudomonas phage PAKP1**
- Pseudomonas phage PAKP1 is a bacteriophage that infects the bacteria *Pseudomonas aeruginosa*
  - This bacteria is the leading cause of healthcare-associated infections in immunocompromised patients
  - Due to increasing resistance towards antibiotics, phage therapy is being studied as an alternative treatment
- This bacteriophage has **relevant clinical applications**, so through this workflow we can study the phage distribution across different environments and its genetic variation

Workflow

This entire workflow was developed and available on Jetstream. https://use.jetstream-cloud.org/application/images/831



Results

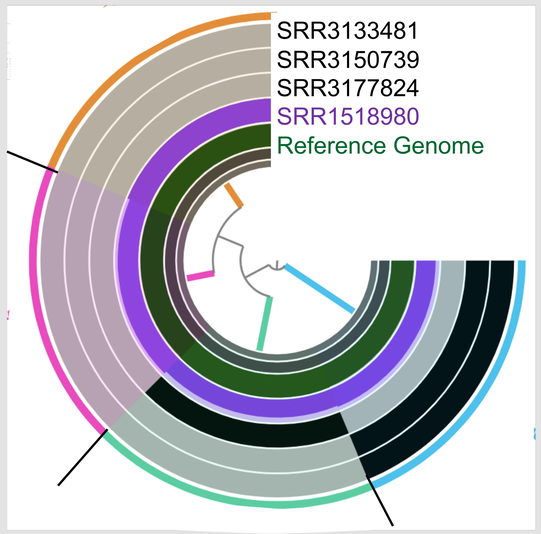


Figure 2: Circular representation of filtered results

- Anvi'o images were generated to show how the results align against Pseudomonas phage PAK1 (Figures 2 and 3)
- Each circle in Figure 2—or horizontal line in Figure 3—represents a dataset, the innermost circle is the reference genome, and the consecutive circles are the metagenomes containing the reference genome (hits from workflow)
- The 4 colored blocks represents a cluster, that was grouped due to sequence similarity. An example of a sequence is shown in Figure 5. A legend shows what the four colored clusters represent (Figure 4).
- The information about these metagenomes are found in Table 1.

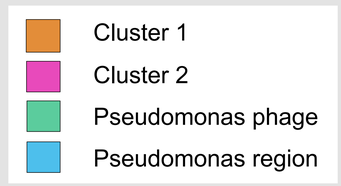


Figure 4: Colored cluster legend

Figure 5: Identified bacterial virus sequence subset

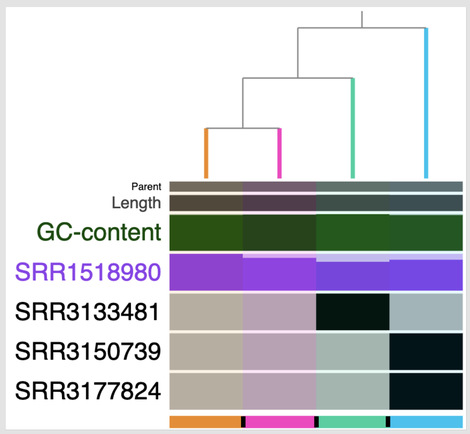


Figure 3: Vertical representation of filtered results

Table 1: Information of the metagenomic datasets identified to contain the phage genome

	SRR3177824	SRR3150739	SRR3133481	SRR1518980
Species Name	Pseudomonas sp. HMSCO71F02	Pseudomonas sp. HMSCO67005	Hydrothermal vent metagenome	Unclassified bacterial virus
Genome Description	HMP Reference genome	HMP Reference genome	Marine sediment metagenome	HMP reference genome
Source	Washington University School of Medicine	Washington University School of Medicine	Zhejiang University	J. Craig Venter Institute
Coverage area on Pseudomonas Genome	**Pseudomonas Region**	**Pseudomonas Region**	**Pseudomonas Phage region**	**Pseudomonas Phage PAK P1 Region**

Results

- Previously classified as an unidentified genome (SRR1518980) in SRA, **it can now be classified as Pseudomonas phage PAKP1**
- Other genomes without full coverage are probably head and/or tail genes, which are shared/conserved between phages
  - **Pseudomonas region** in two samples **(Blue Cluster)**
- Hydrothermal marine vent **(Green Cluster)**
  - BLAST analysis confirmed the sample has elements of Pseudomonas phage and conserved areas of other phages
  - Sequence unique to the **Pseudomonas phage** family
- Other two clusters **(Pink and Orange Clusters)**
  - Unique to the **Pseudomonas phage PAK P1 genome**
  - Had good coverage with reference genome

Discussion

- Why only four Pseudomonas phage PAK P1 genes?
  - Low coverage of searchSRA samples
  - **Too strict with parameters** in workflow
- Future Directions
  - Many genomes on SRA to put through workflow
  - Further analysis on Pseudomonas phage phylogenetics
  - **Test different parameters** for filtering
- Comments
  - Project went in various directions and faced many obstacles which **added valuable feedback** to improving the workflow
  - We tested this pipeline on ZIKV and found metagenome datasets with ZIKA in SRA
  - Pseudomonas Phage PAK P1
    - All SRA, HMP, ambulance, lung metagenomes did not give significant results—surprising!

Conclusion

- We developed a workflow that mines SRA to identify other datasets containing the genome of interest
  - Also worked with crAssphage case study
- Classified an unidentified genome that was on SRA as Pseudomonas phage PAKP1
- Workflow and visualization tool on Jetstream shared out to community for other researchers to use
  - https://github.com/NCGAS/CEWiT-REU-Identifying-datasets-in-SRA-using-Jetstream

Acknowledgements

NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT

Center of Excellence for Women in Technology

Rob Edwards

SDSU